基于科学数据的合作网络研究
——以 ClinicalTrials. gov 临床试验数据为例

徐潇洁[1] 何琳[2] 邵波[1]
[1] 南京大学信息管理学院 南京 210023 [2] 南京农业大学信息科学技术学院 南京 210095

摘要：［目的/意义］基于科学数据构建合作网络，并与传统出版物合作网络进行比较，从网络分析层面解读两个合作网络的差异，为科学数据管理工作提供借鉴。［方法/过程］以 ClinicalTrials. gov 网站的临床科学数据库为例，利用爬虫抓取该网站上传统论文题录信息以及临床试验信息的元数据并分别构建合作网络，通过复杂网络分析比较试验合作机构网络与论文合作机构网络之间的异同。［结果/结论］基于科学数据集和论文数据集的元数据构建的合作网络，与仅从论文数据集中提取元数据构建的网络相比，前者能够展现更丰富准确的合作信息，从而揭示科学数据管理和开放共享的重要性。

关键词：合作网络 科学合作 复杂网络分析 科学数据库 临床试验
分类号：G250
DOI：10. 13266/j. issn. 0252 -3116. 2018. 15. 010

chinaXiv:202308.00608v1

1 引言

科学研究在实验型科研、理论型科研、计算型科研之后已经进入了数据密集型科研的大数据时代。数据是科学发现的基础和驱动，以处理和分析海量数据作为发现知识的基本特征，大数据也被誉为科学研究的“第四范式”。随着 2018 年 1 月 23 日《科学数据管理办法》[1]的全面通过实施，更凸显了数据互通的重要性，从而加快了开放科学数据仓储的发展，科学家们可以不受时间地理的限制进行各方面的合作，基于科学试验数据的新兴合作网络将会受到越来越多的重视。研究科学合作最常用的方式即根据出版物中元数据来提取合作关系，还可以通过问卷调查、定性访问或者三种方法的任意组合来进行探究。但是每一种方法对合作关系的研究都存在一定的局限性，可能会存在高估或者低估的现象[2]。仅仅通过传统论文信息探究合作网络已经不能很好地反映一门学科的发展，而科学数据已经成为非常重要的信息资源，通过分析能从中挖掘到丰富的知识。为适应大数据发展形势，需要加强和规范科学数据管理，同时充分挖掘其中的潜在价值。

本文以 ClinicalTrials. gov 网站的临床试验数据库为例，通过爬虫抓取了该网页的项目研究数据，提取每个研究的传统论文题录信息以及临床试验合作信息的元数据，从而构建科学试验合作网络以及论文合著网络，比较分析他们的异同。

2 相关研究

2.1 科学数据资源库

科学数据资源库如今被使用得非常频繁，但是很少有准确的定义，虽然如此，科学家们对其功能和特征都有一种隐式的共识——即通过收集、注册、观察和创造得出的各种实验数据、观察数据、统计数据等，它可以是论文后附带的实验数据，也可以是独立的研究数据，包括对数据进行描述的元数据、数据集以及数据相关的出版物[3]。它还可以提供额外的数据服务，包括访问、导入、导出、处理、回档以及跟踪和链接到出版物或外部网站等[4]，这些数据是免费的，且在获取、复用上没有知识产权或其他机构的限制，数据的使用完全根据数据拥有者自己的意愿[5]。近几年，开放科学数据得到越来越多的重视，很多国家、机构、大学都在建立开放的科学数据资源库，目的主要是进行数据的复用与共享[6-7]。很多开放科学数据资源库要支持整个领域，所以采用了复杂的技术去运行和维护，这就意味着科学数据仓储的成本高昂，因此这些数据仓储有很

作者简介：徐潇洁（ORCID：0000 -0002 -2851 -3380），硕士研究生，E -mail：574262980@ qq. com；何琳（ORCID：0000 -0002 -4207 -3588）教授，博士；邵波（ORCID：0000 -0002 -6528 -5196）副馆长，教授，博士。
收稿日期：2018 -03 -07 修回日期：2018 -05 -04 本文起止页码：83 -91 本文责任编辑：杜杏叶

ChinaXiv合作期刊

强烈的意愿被使用，在国家政策和法规的支持下被广泛推广，由此科学数据仓储正在影响着科学研究的共享行为，影响着科学合作行为[8-9]。例如，资源型数据库有物理学领域的 LIGO 数据网格，用来支持激光重力波观测试验，约有 500 名科学家参加，其数据对外公开服务。在地球空间科学领域中，美国国家基金会（NSF）和美国国家海洋局（NOAA）资助的 CODIAC 数据库为地球物理研究提供服务[10]。典型的参考型数据库包括蛋白质数据库 PDB、基因序列数据库 GenBank、法国斯特拉斯堡天文数据库 SMBAD、欧洲分子生物学实验室的核苷酸序列数据库 EMBL 等[11]。这些数据资源存储库的使用对科学工作影响的程度以及对科学家们、合作机构合作行为的影响程度都是未知的，在笔者探究这些数据仓储的出现对各个领域科学家合作行为的结构和规模的影响之前，首先要回答更加基本的问题——科学家们在使用这些数据库时进行科学试验合作的结构特点是什么？

2.2 复杂网络分析合作网络

结构和规模的研究涉及到网络中的组成成员、领域内的相互关联、团队的大小等，因此研究合作网络的结构和规模最常采用的方式是复杂网络分析，其中最为著名的就是 R. Albert，A. L. Barabási 和 M. E. J. Newman。A. L. Barabási 等对合著网络随时间演化的现象进行研究[12]，2001 年 M. E. J. Newman 利用社会网络分析法（SNA）对合著网络的研究发现一个科学家只需经过五到六个人就可以与其余任何一个科学家取得联系，科学界似乎形成了一个“小世界”[13]，H. Yang 等[14]发现，个体节点通过和高密度的邻近节点建立联系能够构建一个强强联合的网络。A. Abbasi 等[15]研究了科研合作网络的变化趋势。G. Laudel 将科研合作定义为“一项由多个参与者进行系统合作，以达到研究的目的，从而获得相应的收益”的研究活动[16]，现如今合作已经成为科学生产力发展的主要动力。本文从科学数据资源库中提取两个合作网络，一个是在 ClinicalTrials. gov 网站注册的临床试验机构的合作网络，另一个是基于这些试验发表论文机构的合作网络。

2.3 科学数据集的合作网络研究

2018 年 1 月 23 日《科学数据管理办法》的实施，极大地推动不同领域内科学数据之间、科学数据与其他领域数据之间的流动融合，促进域内专项研究与交叉学科的合作研究。学术界对科学合作进行研究的成果比较多，但是绝大多数限于以科学论文作为研究对象，现在对合作网络的探究开始有了突破，不仅不再基于出版物的元数据，而且开始面向专利，数据仓储等，比如 M. Meyer 和 S. Bhattacharya 首次将专利文献与论文进行比较，虽然两者存在很多不同点，但是在计量上其实有很多相似之处，可以将论文计量的思路运用在专利分析上[17]。J. Singh[18]通过对专利合作网络的探究得出专利合作对于未来信息流动起到推动作用的结论。但是目前，不管是国内还是国际上，关于数据集合作的研究文献非常少，2016 年，为探讨科研合作和大规模的数据仓库在基因组学领域的影响，M. R. Costa[2]基于 GenBank 在大型数据仓库中进行元数据追踪，从传统出版物的合作和数据集的合作中分析合作模式，发现联合分析相关的不同数据集网络能够挖掘出更丰富的信息。陈晓燕[19]构建了 WEB 数据集和论文合著 SCH 数据集并从二值网络和加权网络等角度探讨了不同数据集的异同。由此本文受此启发，将论文计量运用在数据集计量上，基于它们与论文一样，拥有数据持有者，合作者，研究人员等元数据属性。

3 研究方法

3.1 具体方法

本研究运用计量学指标，对 ClinicalTrials. gov 网站注册的临床试验和基于试验发表的论文等情况进行分析；运用 Python 编程完成原始数据向 Netdraw 所需网络文件的转换和基本统计指标的计算；运用 Python 生成合作试验机构与论文合作机构的共现网络文件，并转化为相应的合作网络文件，采用 Ucinet 和 Pajek 软件处理上述合作网络并计算各项指标，其中 Ucinet[20]和 Pajek[21]是综合性基于社会网络的文献信息分析工具，能够支持大型的数据处理，导入的数据需要对其进行加工形成网络格式或者矩阵形式。本研究涉及到的方法包括文献计量法、数理统计法、社会网络分析法（SNA），其中社会网络分析法是基于社会网络中行为者之间关系的量化研究，主要通过点度中心性、中间中心性和接近中心性 3 个指标来进行衡量。

3.2 数据来源

研究科学合作最常用的方式即根据出版物中元数据来提取合作关系，元数据包括作者，机构，期刊，日期等题录信息，基于此来研究合作网络可能会存在高估或者低估的现象。本文选择了 ClinicalTrials. gov 网站作为数据源，从数据库中进行元数据的提取，相比于仅仅提取出版物的元数据，本文的方法使得合作研究更加精确。

ClinicalTrials. gov 网站是全球最大的临床试验登

ChinaXiv:202308.00068v1

记网站,提供了由企业或政府申报的最新的有关临床试验的信息,可通过ClinicalTrials.gov网站检索到全球正在进行的独立开展或国际多中心合作开展的临床试验[22]。本次分析的数据是截止至2016年的全部已完成试验,该网站提供研究目的、研究类型、提交时间、赞助信息、NCT代码、合作机构以及相应的发表的论文等信息。

3.3 数据收集

ClinicalTrials.gov网站允许抓取(https://www.clinicaltrials.gov/robots.txt),而且可以通过适合抓取的网站模式进行综合、无重复抓取(https://www.clinicaltrials.gov/ct2/crawl)。截至2016年12月,该网站共有232 840项临床试验,其中23 551项试验提供了研究结果,209 059项试验没有提供结果的主要原因包括正在进行招募的、试验正在进行中或者由于研究人员的意愿不愿提供。经过清洗(除去重要信息缺失、重复、状态不明确的信息)得到227 503条数据,首先对所有数据进行了描述性分析。笔者又将检索范围定为在ClinicalTrials.gov网站上首次提交时间(First Received Date)为2008年至2016年的全部注册的临床试验,共有182 065条注册信息,对其进行进行清洗、规范化,最终获得164 758条有研究价值的注册信息,其中提供了相应出版物的有45 459条,参与合作(至少与一个机构合作的)的有58 954条。基于出版物数据以及科学数据库之间的差异,本文提出了以下几个问题:①除了在传统出版物的题录信息中提取合作信息,在科学数据库中是否能够提取更加丰富的合作元数据;②基于研究发表的出版物的合作网络与临床试验的合作网络是否存在结构上的差异。针对上述问题,本文通过描述性统计、网络密度、网络平均距离、点度中心性、中介中心性等指标来进行研究。

4 试验项目基本合作情况

4.1 试验合作网络的基本数据

笔者对整理后的数据进行研究与分析,通过统计和计算经过预处理的原始数据集,最终得到了2008－2016年在网站上提交的数据集合作网络的基本数据,这些指标分别是机构数目、连边数目、提交试验数目、平均试验数目、平均度、网络密度、网络直径、平均路径长度以及平均聚类系数。笔者使用Ucinet计算出具体的数值以便于进行分析,详细信息如表1所示:

表1 2008－2016年试验项目合作网络基本信息

	2008	2009	2010	2011	2012	2013	2014	2015	2016	08－16
机构数量	301	366	423	433	492	540	659	748	643	3 484
连边数目	3 708	3 744	4 994	5 228	5 830	6 752	8 738	11 300	8 212	99 012
试验项目数量	1 507	1 529	1 803	1 777	1 902	2 039	2 511	2 653	2 471	18 192
平均项目数	5	4.2	4.3	4.1	3.9	3.8	3.8	3.5	3.8	5.5
网络的平均度	24.64	20.46	23.61	24.15	23.7	25.01	26.52	30.21	25.54	56.84
网络密度	0.041 3	0.028 2	0.028 1	0.028 1	0.024 2	0.023 3	0.020 2	0.020 3	0.02	0.008 2
网络直径	6	8	8	8	9	9	10	9	9	12
平均路径	2.989	3.233	3.417	3.87	3.635	3.602	3.569	3.606	3.803	3.35
平均聚类系数	0.069	0.489	0.486	0.501	0.482	0.502	0.482	0.504	0.507	0.469

考虑到数据规模不能太大而超过所选软件的处理能力,也不能小到无法分析其统计性质,因此本文选取了2008年至2016年期间并且合作机构共现频次大于2的合作数据,其中每组具体的节点数即机构数见表1,每年的节点数均在300到700之间,去掉重复的共有3 484个节点,由于2016年很多项目在招募中,数据不完整,但是通过前几年的趋势能发现每年提交试验的机构数是逐年递增的,临床试验的项目合作网络规模在不断地扩大。相比机构数目,网络的连边数目则跨度较大,在3 000到12 000之间。从表1中看出2008年到2016年网络密度在0.02－0.05之间,可见这9组数据集的网络较为稀疏,团队合作紧密度一般。每年平均路径长度在2到4之间,总体为3.35,这说明临床医学实验领域中,任意一个机构都是可以通过很少的中间人(2－4个左右)到达其他任意一个机构。该网络具有明显的小世界效应,说明该领域信息畅通性强,科研人员合作交流渠道较快捷,信息传播速度较高。但是随着时间的变化,平均路径在不断变长,主要的原因是越来越多的机构参与到合作中来,网络的规模在不断地扩大。10组数据集均具有较低的平均聚类系数,其中2008年的平均聚类系数最低,其他都在0.4到0.5之间,系数很相近,通过观察2008年的机构

chinaXiv:202308.00068v1

ChinaXiv合作期刊

数以及连边数目,平均聚类系数不应该如此低,因此本文通过调查原始数据发现,2008 年的合作的国家跨度很大,合作较为分散,结点的聚类系数普遍低。

4.2 试验合作网络重要属性分析

4.2.1 提交试验数、论文数、平均机构数 对 2008 – 2016 年中至少合作 3 次的项目进行合作规模的统计,不仅统计了每个试验的平均机构数,也统计每篇论文平均机构数,以便于对比发现差异。表 2 列出了每年论文平均机构数和试验平均机构数,每年一篇论文的平均机构数均大于 4,每年一个试验项目的合作机构均大于 3,两个合作网络的变化趋势是类似的,正验证了 M. E. J. Newman 等曾对此问题进行总结,归结其原因是该领域的实验科学研究与理论研究同等重要[23]。但是可以发现篇平均机构数和项目平均机构数呈下降趋势,表明机构之间合作的趋势在不断下降,说明在使用大型或复杂仪器的实验研究(如医学研究领域)中,合作现象是并不是普遍的,这与笔者的预测不一样,因此后面笔者将进行深入的研究并分析其原因。论文网络中一篇论文合作的机构最多达到 16 到 34 之间,而一个试验项目参与合作的机构数最大值在 39 到 92 之间,这说明在试验项目中存在着大范围机构合作研究的现象,而论文的合作范围则较小,这种现象与笔者的直观判断是一致的,由于临床医学试验周期长,所需资源多,其合著范围较大。结合平均机构数的数据可以发现,试验项目合作率虽然不如论文的合作率,但是一旦进行合作,则它的合作规模是很大的,这也是符合实际情况的。

表 2 合作文献(试验)中的平均机构数和最多机构数

年份	论文平均机构数	试验平均机构数	论文最多机构数	试验最多机构数
2008	5.3	5	16	39
2009	5.2	4.2	23	39
2010	5.8	4.3	24	41
2011	5.3	4.1	22	32
2012	4.9	3.9	14	43
2013	4.8	3.8	12	62
2014	4.8	3.8	18	92
2015	4.3	3.5	34	42
2016	4.9	3.8	32	83

4.2.2 度分布 使用 Pajek 软件的 degree 计算功能,得到试验项目合作网络的度分布情况见图 1,度的范围跨越性较大,从 2 到 722,并且主要集中在 50 以下,在该合作网络中,由图 1 可知其分布具有明显的长尾特征。正如 M. E. J. Newman 等的研究发现,科学合作网络更加符合指数截断形式的幂律形式[24],同很多其他的科研合作网络一样具有无标度特性,表明临床医学领域机构试验合作网络会通过增添新节点而继续扩张,而新节点会择优连接到具有大量连接数的节点上,说明少数的机构对于临床医学试验合作网络整体结构形成有重要作用,他们的研究方法、关注焦点的改变会对该领域的发展产生重要影响。比如节点度数最高前三位分别是 Johns Hopkins University(722)、University of California (708)、Massachusetts General Hospital (672),这 3 个机构在这个领域扮演着非常重要的角色。

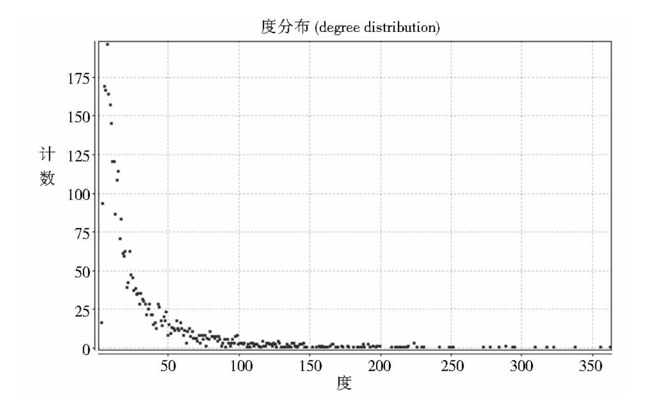


图 1 试验项目合作网络度分布

由此想到基于上述试验而发表的论文合作网络的度分布是否与其存在差异,笔者同样将论文的合作机构的数据导入 Pajek 软件中,得到论文合作网络的度分布见图 2,可以发现横坐标与纵坐标的刻度明显变小,说明论文合作网络规模比试验项目合作网络的规模要小,度分布的跨度为 2 到 184,远小于试验合作网络,但它同样具有长尾特征,同样符合幂律分布,同样具有无标度特性,节点度数最高前三位分别是 Columbia University(184)、Massachusetts General Hospital(181)、University of California(177),论文合作网络和试验合作网络并不是一样的,论文合作较多的机构或者说在临床医学这个领域内学术成就处于中心地位的机构并不一定在试验项目合作中占据重要地位。同时,临床试验合作网络节点度数跨度很大,最大度数达到 722 之间,分化和抱团现象严重,存在着对临床医学领域影响力较大的机构并且经常性的进行合作;而论文对应的节点度较低,节点数对应最多的 2,说明论文网络相对于试验项目网络不会存在严重的抱团现象,合作的可能性和选择比较多。

chinaXiv:202308.00508v1

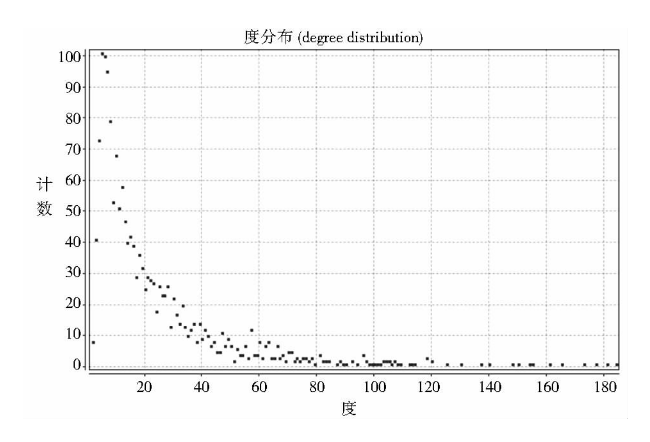


图 2　论文合作网络度分布

5　试验项目合作与其发表论文合作情况

5.1　项目发表论文情况分析

科学论文更是衡量“科学生产力”的重要指标[25]，在 ClinicalTrials. gov 数据库中除了提供相关的实验项目信息，还提供了基于该项目所发表的论文信息，因此本文按照发表论文数进行递增排列，探究临床医学试验中发表论文的能力。从表 3 中可以发现，有 162 512 个项目数并没有发表相应的论文，占据了全部项目数的 69. 8% 左右，这是相当大的比例，并且大部分项目发表的论文数集中在 1－10 篇，其中有很多合作规模较大的试验项目发表的论文不多甚至没有发表论文，例如 Pfizer（目前全球最大的以研发为基础的生物制药公司）在 ClinicalTrials. gov 中提交了 1 174 个试验项目，其中有 803 个实验项目没有发表任何论文，约占了 68. 4% 左右，甚至有很多与多个机构进行合作的项目也没有提供任何论文信息，是否这部分试验发表了专利而不是论文？经过网站调查，发现 Pfizer 和 Bristol-Myers Squibb 在网上合作的试验有 19 个，其中有 14 个都是基于 Apixaban 的临床试验研究，但是这 14 个项目中只有 2 个项目发表了论文，通过专利调查，发现 Pfizer 和 Bristol-Myers Squibb 在欧洲专利数据库中检索到其在 2016 年 10 月 25 日发表名为 APIXABAN FORMULATIONS 的专利，正是在试验提交时间的区间内。可见如果仅仅以论文合作情况作为元数据探究合作网络明显受到了制约。这也回答了为什么上文探究提交试验数以及论文数的平均机构数时发现，合作率在下降，这很有可能是因为发表了相关专利，尤其是在临床医学这个非常重视专利和知识产权的领域。从另一角度还可以发现有部分项目发表的论文数非常多，如发表了 130 篇以上的项目数就有 12 个，平均一个项目发表约为 11 篇论文，可见还有项目带来的论文产量是如此之多，更加说明仅仅从论文的角度来衡量一个机构的科研能力和合作特点是以偏概全的。

表 3　不同论文数量的项目数分布

论文数	项目数	论文数	项目数	论文数	项目数
0	162 515	14	516	28	97
1	30 213	15	416	29	87
2	5	16	379	30	66
3	8 616	17	360	31－40	610
4	7 285	18	284	41－50	277
5	3 634	19	264	51－60	140
6	2 702	20	221	61－70	78
7	1 896	21	217	71－80	49
8	1 445	22	176	81－90	23
9	1 094	23	172	91－100	12
10	992	24	163	101－110	19
11	832	25	140	111－120	12
12	643	26	124	121－130	3
13	581	27	135	>130	12

5.2　合作网络对比分析

上述研究发现，仅仅通过研究论文的合作网络是很难精确探究科学合作网络的，因为很多试验项目由于商业性质，政府政策，个人因素等并没有发表相应的论文，尤其是临床医学这类领域，很有可能会以专利的形式展现试验成果，反过来有些机构一段时间内发表的合作论文很可能仅基于一次试验项目的合作。因此试验项目的合作网络和基于这些项目发表论文的合作网络应当存在异同。下面将从网络密度，平均路径长度、中心性等角度进行分析。

5.2.1　合作密度分析　2008－2016 年中有 8 275 个机构共进行了 180 139 项试验（删除了由于合作、转载、更正等造成的重复现象以及没有提供任何信息状态的现象），为了保证数据的准确性，删除了其中权重小于 10 的边，以及只有孤立的结点，一定程度上剔除了合作的随机性和偶然性，最终得到了 177 个结点，564 条边的合作网络，见图 3。

经计算，该网络密度为 0. 0210，聚类系数为 0. 542，该网络呈现的是以部分度值较高的研究机构为局部中心点的紧密的、大范围和小范围都存在的合作关系，说明这些机构凝聚程度很高，发生合作的可能性很大，知识整合广度高。此外，与该核心网络相连通的节点还形成了以 Johns Hopkins University、University of Washington、National Cancer Institute（NCI）等为局部中心点的合作小网络，同时网络周围还存在较多散在的合作对。

chinaXiv:202308.00008v1

ChinaXiv合作期刊

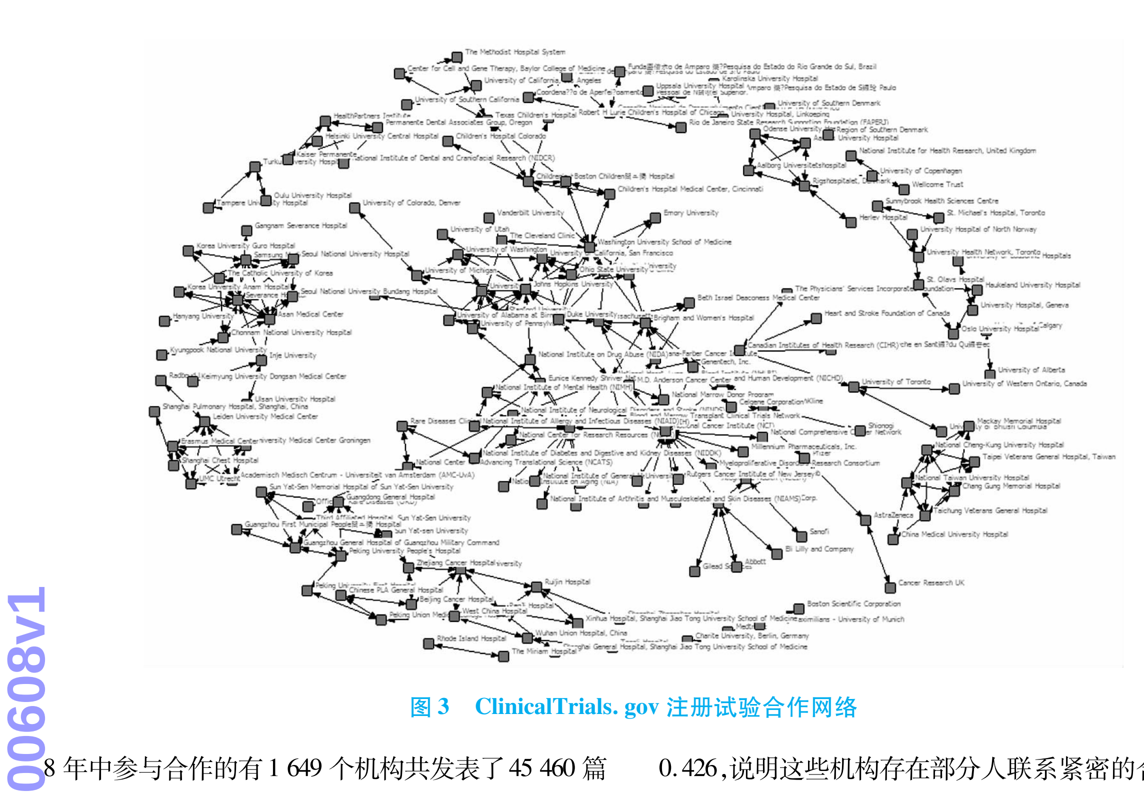


图 3 ClinicalTrials. gov 注册试验合作网络

8 年中参与合作的有 1 649 个机构共发表了 45 460 篇论文,同样笔者删除合作次数小于 10 的边,以及孤立的结点,构建了包含 25 个结点,48 条边的合作网络,见图 4。经计算,该网络密度为 0. 0104,密度较小,聚类系数为 0. 426,说明这些机构存在部分人联系紧密的合作团队,但相互之间缺乏广泛合作,知识整合广度和知识整合效率不高,合作模式单一,缺乏能够连接不同合著群体桥梁的作者,缺乏知识创新速度和可持续性的有力保障。

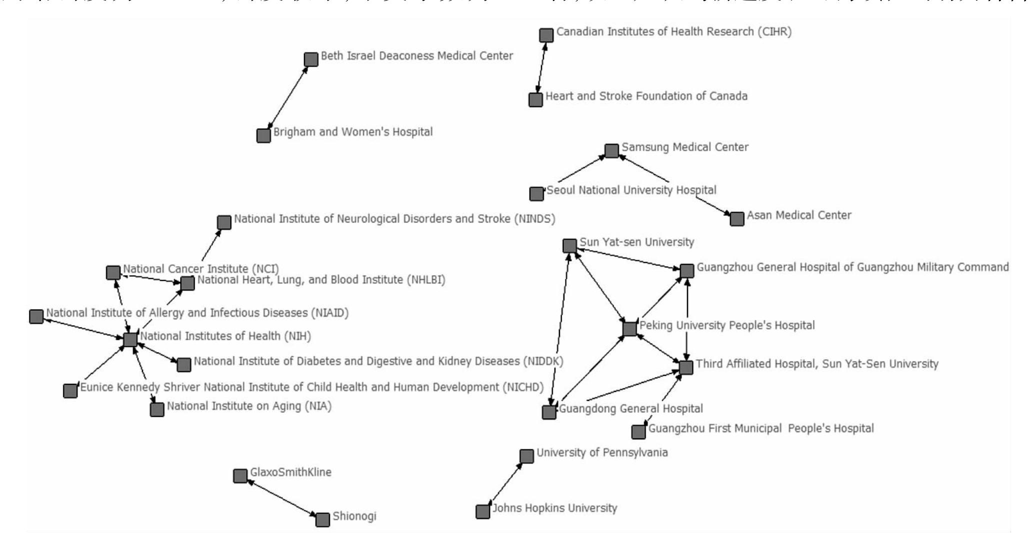


图 4 ClinicalTrials. gov 论文合作网络

5.2.2 典型小团体网络分析 对这些网络进行局部研究,在论文合作网络中,两两机构合作的情况最多(更多的是机构内部的合作,这里不予考虑),其中合作最多的是 CNPq 和 Fundação de Amparo à Pesquisa 形成的科研团队。在试验项目合作中,通过 k－核的分析,一些紧密团结在一起的群体,其中最大的一个群体是如图 5 中的 3 个, k 值为 8,说明每个机构都至少与其他 8 人产生合作关系。第一大子群中 9 个机构的颜色一样,并没有特别明显的核心人物,但频繁进行合作。网络中第二大和第三大子群分别以 UMC Utrecht(荷兰最大的大学医疗中心之一)和 Seoul national University Hospital(首尔大学)为核心的合作网络,他们两个机构是两个子群交流的主要枢纽。论文合作网络主要以两两合作现象为主,可探究的信息相比试验项目合作网络要少很多,足以发现试验合作网络对一个学科发展的意义。

chinaXiv:202308.00608v1

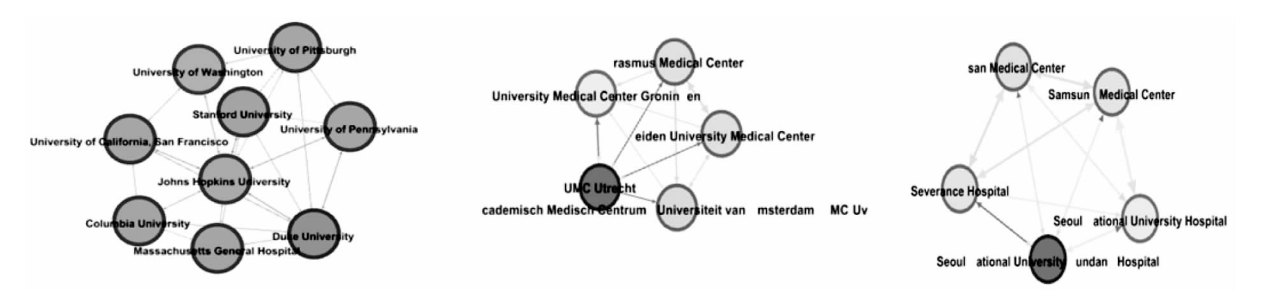


图5 k核值为8的三个合作小团体

5.3 合作网络中心性分析

5.3.1 点度中心性 点度中心性分为两类:绝对中心度和相对中心度,其中相对中心度则是比较不同网络的结点中心性的指标[26]。利用Ucinet的Degree算法对两个网络进行分析,见表4(表中仅列出排名前10的机构)。从结果中笔者可以看出,试验网络点度中心度最高的是Johns Hopkins University,其绝对点度中心度是360,表明Johns Hopkins University曾与360个机构有过合作试验,其知识扩散能力强。可以看出Johns Hopkins University在临床试验的网络中的地位很高,很有可能由于其科研能力、设备水平、科研人员水平而使得其他机构愿意与之合作,其相对中心度为10.336,而在论文网络中Beijing Chao Yang Hospital的点度中心度最高,其相对点度中心度为5.522,前者远远高于后者,说明前者对试验合作网的支配能力可能大于后者对论文合作网的支配能力。合作试验多的机构并不是合作发文多的,论文点度中心度和试验点度中心度不呈现显著相关($r=0.479$, $P>0.05$),也就是说,合作论文的机构数量与合作试验的机构数量不相关。

表4 合作试验(文献)点度中心度

排名	试验机构	绝对	相对	论文机构	绝对	相对
1	Johns Hopkins University	360.00	10.336	Beijing Chao Yang Hospital	91.00	5.522
2	University of California, San Francisco	353.00	10.135	Hospital Universitario Ramon y Cajal	88.00	5.340
3	Massachusetts General Hospital	335.00	9.618	St. Joseph's Hospital and Medical Center, Phoenix	87.00	5.279
4	National Cancer Institute (NCI)	320.00	9.187	Harvard University	87.00	5.279
5	Columbia University	315.00	9.044	DokuzEylul University	86.00	5.218
6	Duke University	307.00	8.814	Ministry of Health, Spain	85.00	5.158
7	National Institutes of Health (NIH)	292.00	8.384	Shionogi	85.00	5.158
8	Stanford University	291.00	8.355	Aurora Health Care	85.00	5.158
9	Mayo Clinic	286.00	8.211	Flevoziekenhuis	83.00	5.036
10	University of Michigan	286.00	8.211	Seventh Framework Programme	79.00	4.794

5.3.2 中间中心性 中间中心性用来测量的是机构对资源掌控的程度。对论文网络和试验网络进行中间中心性分析,如表5所示:

表5 合作试验(文献)中间中心度

排名	试验机构	绝对	相对	论文机构	绝对	相对
1	Fudan University	193 788.203	3.196	Flevoziekenhuis	203.229	0.015
2	Pfizer	183 065.969	3.019	St. Joseph's Hospital and Medical Center, Phoenix	188.352	0.014
3	KarolinskaInstitutet	161 705.375	2.667	Triemli Hospital	166.654	0.012
4	National Cancer Institute (NCI)	161 601.656	2.665	Kangdong Sacred Heart Hospital	164.038	0.012
5	University of California, San Francisco	148 560.047	2.450	Naval Medical Research Center	152.995	.011
6	Canadian Institutes of Health Research (CIHR)	142 747.297	2.354	Beijing Chao Yang Hospital	151.954	0.011
7	Johns Hopkins University	134 332.203	2.215	Hospital Universitario Romany Cajal	151.576	0.010
8	Merck Sharp &Dohme Corp.	130 519.664	2.152	Catholic University, Italy	141.704	0.010
9	GlaxoSmithKline	124 497.781	2.053	University of Cologne	140.203	0.010
10	Massachusetts General Hospital	119 528.078	1.971	Daegu Catholic University Medical Center	138.215	0.010

chinaXiv:202308.00081

ChinaXiv合作期刊

首先是试验机构网络的中介中心度,最高的是 Fudan University,其次是 Pfizer、Karolinska Institute、National Cancer Institute(NCI)等机构,这些机构的中介中心度比较高,说明他们掌握了很多的研究资源,有一部分机构的中介中心度接近于 0,共有 302 个,这些机构不具备控制资源的能力,占机构总数 9.12%。论文机构网络中介中心度,最高的是 Flevoziekenhuis,St. Joseph's Hospital and Medical Center 和 Phoenix Triemli Hospital 紧随其后,两极分化现象严重,其中中介中心度为 0 的有 1 514 个,占总数 1 674 个的 90.44%,少于 10 的占 91.63%。可以明显发现实验机构网络中中间中心性为 0 的机构比论文机构网络中要少很多,说明在论文网络中影响力较强的"中间人"机构较少,对网络中其他人的合作关系控制力强度不够,出现这一结果的原因,可能与临床医学信息学领域作者合著交流规模不广泛有关,也同样反映出专利存在的可能性。

6 结论

本研究通过对被 ClinicalTrials. gov 网站收录临床试验数据的合作情况进行文献计量和网络分析,在探究试验项目合作情况的同时,对比分析了基于试验发表论文的情况,得出论文合作情况和数据集合作情况在合作文献量、合作率等计量测度以及密度、中心性分析、平均路径等网络测度上的异同,揭示出两种合作网络的区别和内部联系,从而为后续的科学合作和技术合作研究提供独特视角。主要结论如下:

(1)项目合作总体情况。2008 年 –2016 年期间,临床医学领域内机构的合作规模以及平均每个机构所做的项目数呈现上升的趋势,合作的密度也在逐年增加,说明在该领域数据集的合作意识在不断加强,但是合作的广度欠缺,具有明显的抱团现象,主要以几个合作机构为中心进行合作,节点度数最多的机构并不是中心度最高的机构,说明合作次数最多的机构只与固定的几个机构进行合作,而不是进行广泛的合作。

(2)基于论文合作网络对试验项目合作网络的分析。现如今"第四范式"的发展正在影响着科学合作的规模与结构,也在支持着跨国跨领域跨机构的合作。本文提出科学数据的元数据比仅使用出版物元数据能提供更多关于科学家合作行为信息的猜想,由上文可以看出,这个假设是成立,例如基于中心性的分析,在两个网络中点度中心性和中间中心性排名前十的机构是不一样的,甚至差别很大,合作项目最多的团体并不是合作论文最多的团体,如果仅仅通过论文来探究合作特征,就会出现很大的误差,尤其很多机构在试验项目中进行了紧密的合作,但是并没有发表相应的论文,可能选择了发表相关的专利,这种情况在临床试验领域中非常普遍。同时,科学家相对于首先在出版物上合作,更可能在正式出版合作之前就在数据集上合作,毕竟产生的数据只是研究过程中的一部分,发表出版物才是最后的目的,因此往往实力较弱的机构更倾向于试验的合作。

论文和科学数据作为科学研究的两大产出,在计量学中,对论文和数据集的研究是平行的、可比的。目前,情报学界对基于科学数据集探究合作网络方面的研究还有待加强,本研究从网络分析层面初步解读了论文合作网和数据集合作网的差异,是论文和数据集比较研究的尝试,也为该研究提供一个新的视角,也揭示了科学数据的重要性,有必要加强和规范科学数据管理以适应大数据的发展形势。

参考文献:

[1] 思想再解放改革再深入工作再抓实 推动全面深化改革在新起点上实现新突破[N]. 人民日报,2018-01-24(1).

[2] COSTA M R, QIN J, BRATT S. Emergence of collaboration networks around large scale data repositories: a study of the genomics community using GenBank[J]. Scientometrics,2016,108(1):21-40.

[3] 黄永文,张建勇,黄金霞,等. 国外开放科学数据研究综述[J]. 现代图书情报技术,2013(5):21-27.

[4] MARCIAL L H, HEMMINGER B M. Scientific data repositories on the Web: An initial survey[J]. Journal of the Association for Information Science and Technology,2010,61(10):2029-2048.

[5] 张晓林. 开放获取、开放知识、开放创新推动开放知识服务模式——3O 会聚与研究图书馆范式再转变[J]. 现代图书情报技术,2013(2):1-10.

[6] PILLAI B. Cyberinfrastructure essential to 21st century advances in science and engineering education & research[C]//International conference on control, automation and systems. Seoul: IEEE, 2007:71-73.

[7] HEY T. The fourth paradigm - data-intensive scientific discovery[J]. Proceedings of the IEEE, 2011, 99(8):1334-1337.

[8] FANIEL I M, JACOBSEN T E. Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues' data[J]. Computer supported cooperative work, 2010, 19(3):355-375.

[9] FANIEL I M, ZIMMERMAN A. Beyond the data deluge: a research agenda for large-scale data sharing and reuse[J]. International journal of digital curation,2011,6(1):58-69.

[10] 傅小锋,李俊,黎建辉. 国际科学数据的发展与共享[J]. 中国基础科学,2007,9(2):30-35.

chinaXiv:202308.00068v1

[11] 刘闯,孙鸿烈. 国际科学技术数据前沿领域发展研究[J]. 中国基础科学,2003, 18(1):329 - 333.
[12] BARABASI A L, JEONG H, NEDA Z, et al. Evolution of the social network of scientific collaborations[J]. Physica a: statistical mechanics and its Applications, 2001,311(3/4):590 - 614.
[13] NEWMAN M E J. Erratum: scientific collaboration networks. II. Shortest paths, weighted networks, and centrality[J]. Physical review e statistical physics plasmas fluids & related interdisciplinary topics,2006,73(3):039906.
[14] YANG H, WANG W, WU Z. Diversity-optimized cooperation on complex networks[J]. Physical review e,2009,79(5):56107.
[15] ABBASI A, HOSSAIN L, LEYDESDORFF L. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks[J]. Journal of informetrics,2012,6(3):403 - 412.
[16] GRIT L. What do we measure by co-authorships? [J]. Research evaluation,2002,11(1):3 - 15.
[17] MEYER M, BHATTACHARYA S. Commonalities and differences between scholarly and technical collaboration-an exploration of co-invention and co-authorship analyses[J]. Scientometrics,2004,61(3):443 - 456.
[18] SINGH J. Collaborative networks as determinants of knowledge diffusion patterns[J]. Management science,2005,51(5):756 - 770.
[19] 陈晓燕. 学术数据集和 WEB 数据集下著者社会网络的比较研究[J]. 情报科学, 2014(5):79 - 84.
[20] Pajek[EB/OL]. [2017 - 04 - 15]. http://www.pajek.imfm.si/doku.php? id = pajek.
[21] UCINET[EB/OL]. [2017 - 04 - 15]. http://www.analytictech.com/ucinet/.
[22] THELWALL M, KOUSHA K. Are citations from clinical trials evidence of higher impact research? An analysis of ClinicalTrials. gov[J]. Scie-ntometrics,2016,109:1 - 11.
[23] NEWMAN M E J. The structure of scientific collaboration networks[J]. Proceedings of the national academy of sciences of the United States of America,2001,98(2):404 - 409.
[24] NEWMAN M E J. The structure and function of complex networks[J]. 2003,45(2):167 - 256.
[25] 王文军,袁翀. 社会科学学术论文生产力评价的新视角——C100 指数的理念、构建方法及其初步测试[J]. 山东社会科学,2015(2):186 - 192.
[26] 邱均平,瞿辉. 我国科研机构合作网络知识扩散研究——以"生物多样性"研究为例[J]. 图书情报知识,2011(6):5 - 11.

作者贡献说明:
徐潇洁:论文框架设计,论文撰写与修改;
何琳:论文选题,提出论文修改建议;
邵波:论文选题与设计,提出论文修改建议,论文定稿。

Research on Collaboration Network of Scientific Data-Taking Clinical Trial Data of ClinicalTrials. gov as an Example

Xu Xiaojie[1] He Lin[2] Shao Bo[1]
[1] School of Information Management, Nanjing University, Nanjing 210023
[2] School of Information Management, Nanjing Agriculture University, Nanjing 210095

Abstract: [**Purpose/significance**] Based on the scientific data, we constructed a new collaboration network and compared with cooperative network of traditional publications. By exploring the differences of the two collaboration network from the perspective of network analysis, this paper provided reference for scientific data management. [**Method/process**] Taking clinical science database of ClinicalTrials. gov website as an example, this paper used crawler technology to capture the metadata of traditional publication and clinical trial research. Then, based on these two kinds of metadata, this paper constructed different cooperative networks respectively. Finally, it used complex network analysis to explore these networks to compare the similarities and differences between two networks. [**Result/conclusion**] Scientific collaboration network extracted metadata from a scientific repository and publication could provide richer and more accurate information of collaboration than just using metadata from publications alone. This paper reveals the importance of scientific data management and open sharing.

Keywords: collaboration network scientific collaboration complex network analysis scientific database clinical trial

chinaXiv:202303.00608v1